**COVID-19 Data Analysis**

Siddharth Dudugu, Harsh Ramesh, Priyanka Choudhary

**Background:**

The COVID-19 pandemic has brought unprecedented challenges to global healthcare systems, emphasizing the critical need for efficient resource allocation. Throughout the course of the pandemic, healthcare providers have grappled with shortages of medical resources, highlighting the necessity for predictive tools to anticipate and address the varying needs of COVID-19 patients. This project focuses on leveraging Apache Spark for data analysis to develop a machine learning model capable of predicting the risk level of COVID-19 patients. By analyzing patient symptoms, health status, and medical history, the model aims to assist healthcare authorities in allocating resources effectively and prioritizing high-risk cases, contributing to timely interventions and improved patient outcomes.

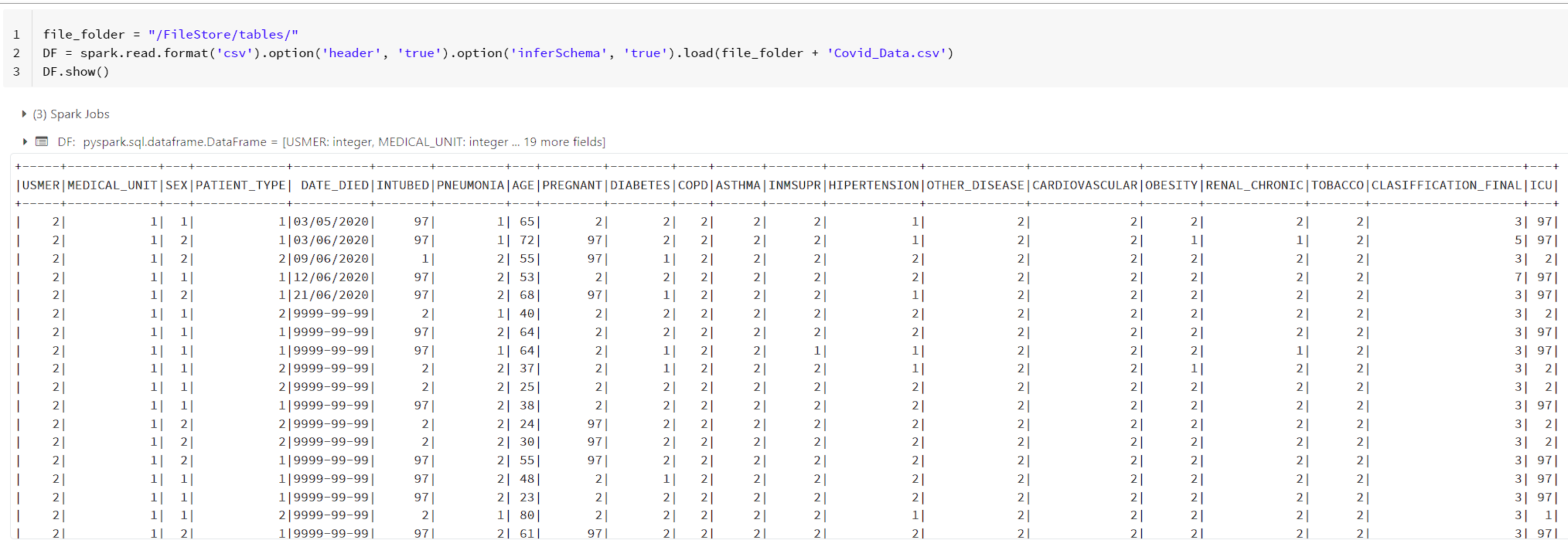
**Dataset:**

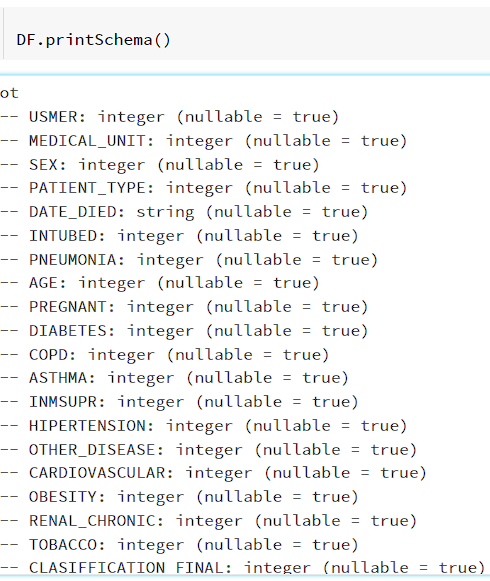
**Link:** <https://www.kaggle.com/datasets/meirnizri/covid19-dataset>

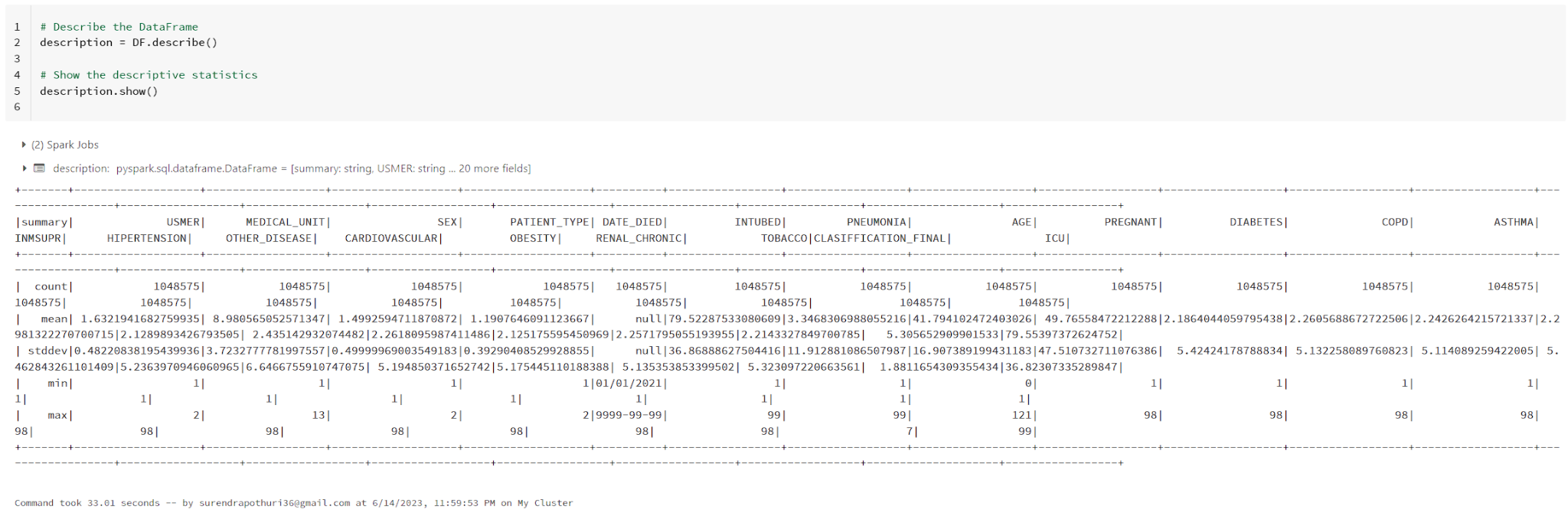
The dataset was provided by the Mexican government (link). This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data.

* sex: 1 for female and 2 for male.
* age: of the patient.
* classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different
* degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
* patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
* pneumonia: whether the patient already have air sacs inflammation or not.
* pregnancy: whether the patient is pregnant or not.
* diabetes: whether the patient has diabetes or not.
* copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
* asthma: whether the patient has asthma or not.
* inmsupr: whether the patient is immunosuppressed or not.
* hypertension: whether the patient has hypertension or not.
* cardiovascular: whether the patient has heart or blood vessels related disease.
* renal chronic: whether the patient has chronic renal disease or not.
* other disease: whether the patient has other disease or not.
* obesity: whether the patient is obese or not.
* tobacco: whether the patient is a tobacco user.
* usmr: Indicates whether the patient treated medical units of the first, second or third level.
* medical unit: type of institution of the National Health System that provided the care.
* intubed: whether the patient was connected to the ventilator.
* icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
* date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

**EDA:**







A diagram of a number of patients

Description automatically generated A graph of age distribution

Description automatically generated

A graph with blue and orange bars

Description automatically generated

**Project Goals:**

The project aims to utilize logistic regression and other relevant machine learning algorithms to develop a robust predictive model for identifying high-risk COVID-19 patients. By employing feature selection techniques and leveraging a preprocessed dataset containing critical patient information, the goal is to accurately assess the model's performance using metrics such as accuracy, precision, recall, and F1 score.